Abstract: MicroRNAs, or miRNAs, are small single-stranded non-coding RNAs noted for their involvement in gene silencing pathways. Some miRNAs have target genes implicated in human disease, such as oncogenes and tumor suppressors, and many studies have now shown that alterations to miRNAs (i.e. their mutation or deletion) play a role in disease progression[1]. Additionally, epitranscriptomics, or the investigation of chemical modifications on nucleic acid single bases, has recently been proposed as a potential source of biomarkers for early diagnostics. The epitranscriptome of many RNA species have been characterized; however, whether or not these modifications are present in miRNAs remains unknown. Here, we present an investigation of the propensity of base modifications in miRNAs in human cell lines using RNA-seq and mass spectrometry. Specifically, we introduce a robust characterization of the modifications found on miRNAs released extracellularly in exosomes, which are a particularly useful biomarker since they are found in the bloodstream rather than in tissues. Our findings reveal that the representation of RNA species in exosomes deviates from that of total cell contents. We also show evidence of novel modifications in exosomally-released miRNAs.

Introduction: Exosomes are a type of extracellular vesicle (EV) about 30-100 nm in size which originate in cells and are plentiful in biological fluids[2]. Previous studies have shown that these vesicles contain RNA species, known as exoRNA. The abundance of exosomes in biological fluids, as well as the non-invasive nature of biological fluid sample collection, gives them the potential to serve as superior biomarkers for precision medicine. Exosomes have also recently been shown to contain miRNAs, termed exomiRs, and there is increasing evidence that the movement of miRNAs into exosomes

is non-random[3]. Therefore, exosomes remain viable and advantageous candidates for biomarkers.

MicroRNAs are small non-coding RNAs that have important functions in the cell proliferation cycle and cell death[4]. Since errors in the cell cycle can lead to uncontrolled cell growth and tumor formation, alterations to microRNAs or their control were thought to play a role in cancer progression. Many studies have now shown this to be true[1]. These studies have typically been concerned with the mutation, amplification, or deletion of the genes encoding the microRNAs.

However, an alternative mechanism for regulating RNA function is the base methylation of RNA itself. This is a phenomenon which has been studied previously in other types of RNA, such as tRNA and mRNA, but not in microRNAs[5]. It is likely that methylation is another way that cancer cells misregulate microRNAs. Two such modifications that are particularly interesting are methyl ^{1}A and ^{1}G, since these have been shown to affect Watson-Crick base pairing in nucleotide base pairs[6]. This is a particularly notable phenomenon in miRNAs since one of their main functions is to repress target genes by Watson-Crick base pairing with a target mRNA. This suggests that there would be a profound difference in functionality between methylated and non-methylated miRNAs, which could greatly influence regulatory mechanisms in the cell cycle.

Here, we aim to characterize the epitranscriptomic profile of miRNAs isolated from total cell contents as well as exosomes. We compare the abundance of miRNAs found in each of these conditions to test the hypothesis that their sorting into exosomes is non-random. Additionally, by probing the modification landscape of small RNAs isolated under each of these conditions, we identify novel exomiR modifications that yield insight into the functions of exosomal RNAs.

Results:

A variety of methods are commonly used to probe the modification landscape of nucleic acids. Liquid chromatography-mass spectrometry is often used to distinguish modifications based on size. For the purpose of these experiments, LC-MS was run in multiple reaction monitoring (MRM) mode. MRM mode first selects for precursor ions of a defined size, fragments these, and then selects for fragmented products of a second defined size. MRM mode is often run to detect known mass transitions of common base modifications in order to reduce noise.

Once LC-MS results are analyzed and abundant modifications in the sample of interest are identified, tandem antibody pull down RNA-seq can be used to identify the species containing the modification of interest. Antibodies that select for modifications of interest are used to isolate all RNA species containing the modification. cDNA libraries can then be prepared from the isolated RNA, and subsequent sequencing reveals a comprehensive list of all RNA species containing that modification. For the purposes of the RNA-seq experiments presented here, antibody pull downs were not used.

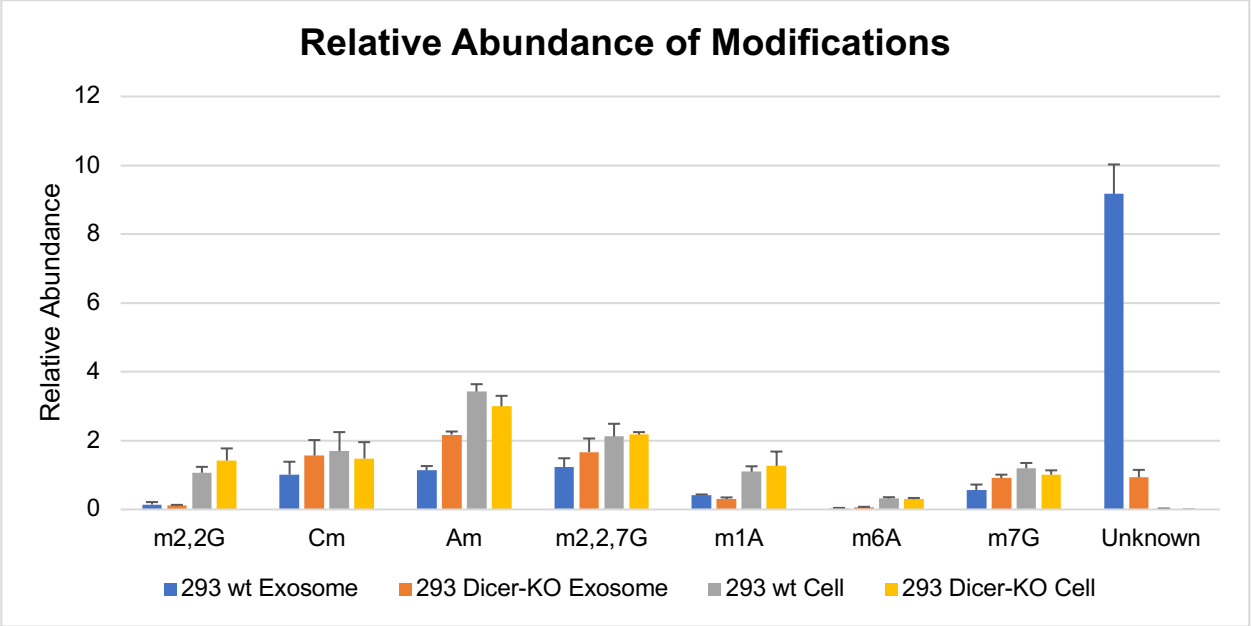


Fig 1: The relative abundance of base modifications found in HEK293 wild-type cells and exosomes, and HEK293 Dicer-knockout cells and exosomes. RNAs isolated from the exosomes and total cell contents of wild-type HEK293 and Dicer-knockout HEK293 cells were hydrolyzed into single-nucleotide fragments analyzed using mass spectrometry. Total unprocessed spectra are not shown. The abundance of a number of base modifications was determined in a series of three replicate experiments. The average of these biological replicates is shown here. No attempt was made to separate RNA species based on size, and this graph therefore represents a multitude of RNA species.

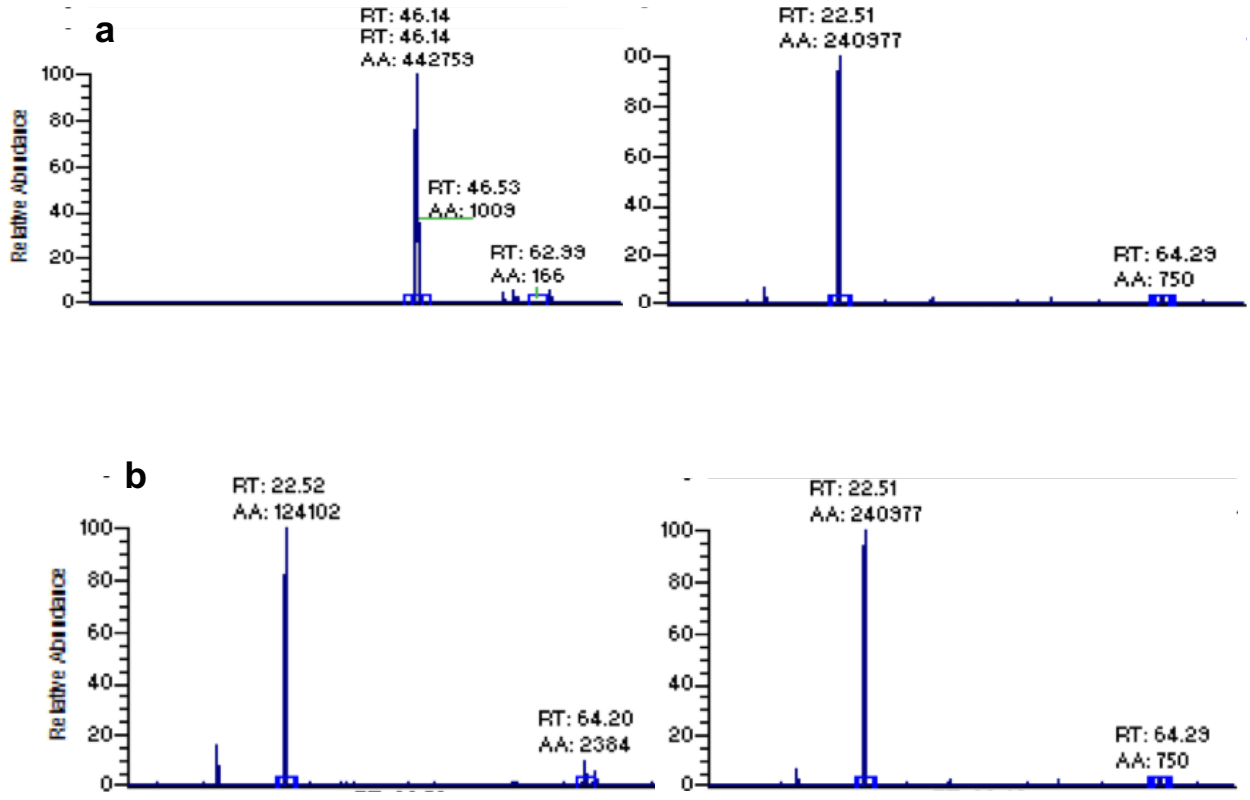


Fig. 2: Mass spectrometry data confirms that the 299.1 amu peak preferentially found in wild-type exosomes is not N4-acetyl-2′-O-methylcytidine. (a) The spectrum on the left is an ac4Cm standard, and the spectrum on the right is RNA isolated from HEK293 wild-type exosomes. The vastly different retention times of the two peaks (46 minutes versus 22 minutes, respectively) indicates that the experimental peak of interest is not ac4Cm. (b) The spectrum on the left is small RNAs (15-35 nt) and the spectrum on the right is long RNAs (50-75 nt) isolated from HEK293 wild-type exosomes. The 299.1 amu peak at 22 minutes appears at equal abundance in both samples.

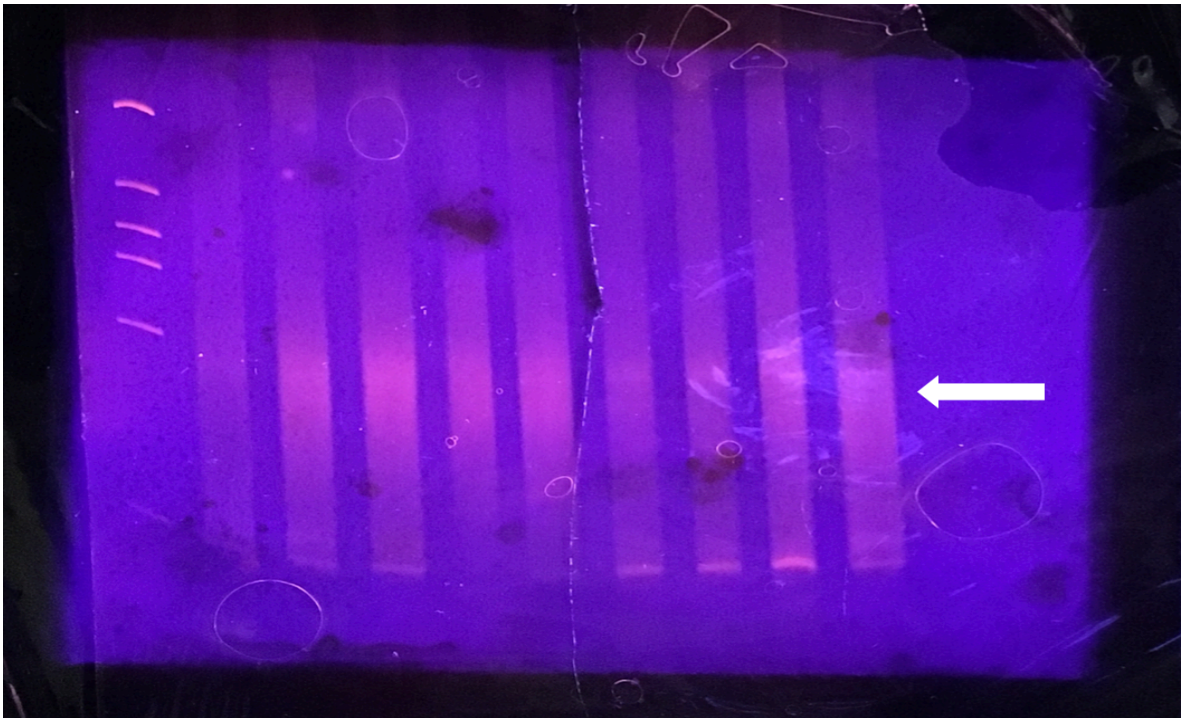


Fig. 3: Polyacrylamide gel showing the band of rRNA-depleted small RNAs for downstream purification and sequencing. A fraction of RNAs isolated from human HEK293 wt, HEK293 Dicer-KO, and BCB1 cells and exosomes were fragmented to control for the bias toward small RNAs in RNA-sequencing (see Materials and Methods). Additionally, the fragmented total cellular RNA was rRNA depleted to maximize the mapping rate of small RNAs during RNA-seq. Here, the bright bands indicated with a white arrow represent cDNA libraries approximately 120 – 160 nt in length, which were gel extracted and sent for sequencing.

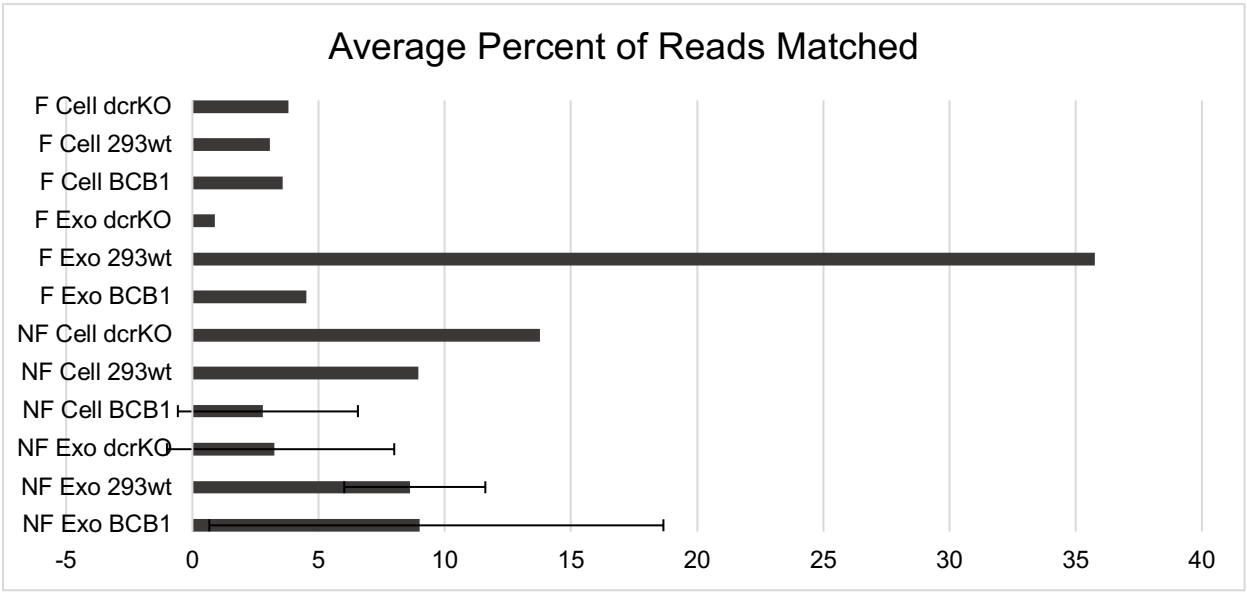


Fig. 4: Average read mapping rate of replicate RNA-seq experiments is exceedingly low. Small RNA libraries were created from both fragmented and non-fragmented samples, which were then sent for RNA sequencing. The reads of these sequencing runs, performed in triplicates, were matched to a database of known miRNAs; however, the mapping rate was very low compared to the typical mapping rates of these experiments.

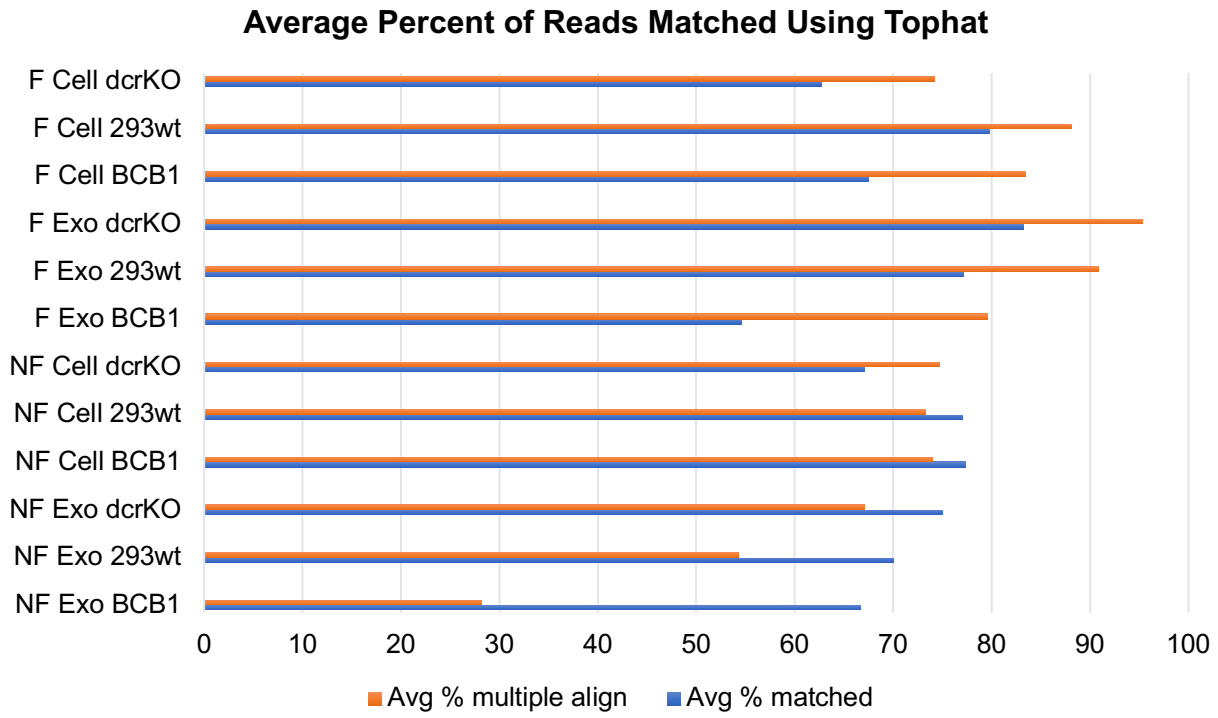


Fig. 5: Average read mapping rate and multiple alignment rate of replicate RNA-seq experiments reveal sequencing issues. Small RNA libraries were created from both fragmented and non-fragmented samples, which were then sent for RNA sequencing. The reads were assembled to the human genome using the TopHat command-line utility, which maps across splice junctions[7]. The mapping rate increased significantly (blue bars), however, many of the reads mapped multiple times in the genome (orange bars), suggesting that the mapping rate may be artificially inflated.

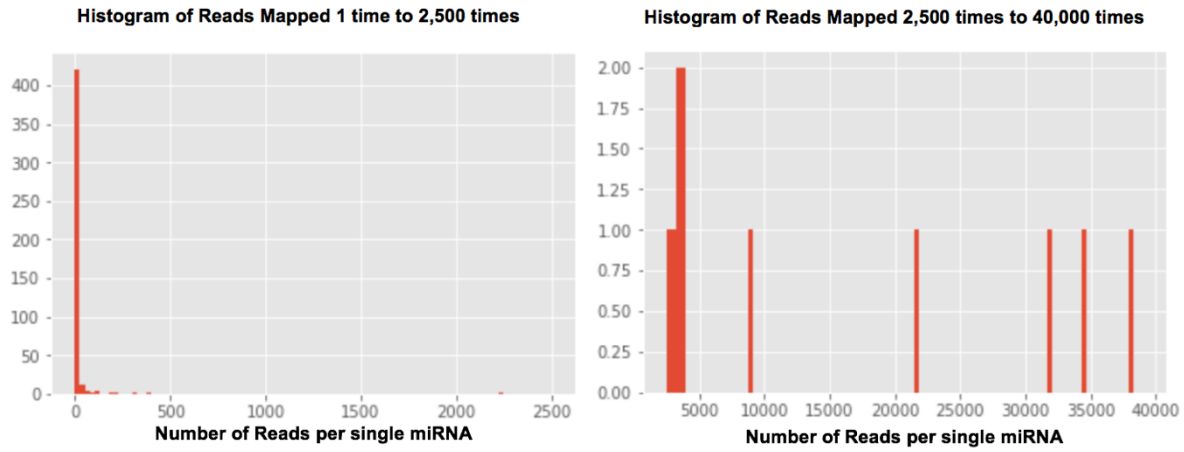


Fig. 6: Histograms of read mapping reveal contamination in sequencing samples. The left panel shows a high number of miRNA species (400+) that were hit between 1 and 500 times, while the right panel shows a low number of miRNA species (8 shown) that were hit an extremely high number of times. This is indicative of contamination of these 8 species. Contaminant species were removed prior to downstream analysis using a length filter (see Materials and Methods). Plots were generated using matplotlib.

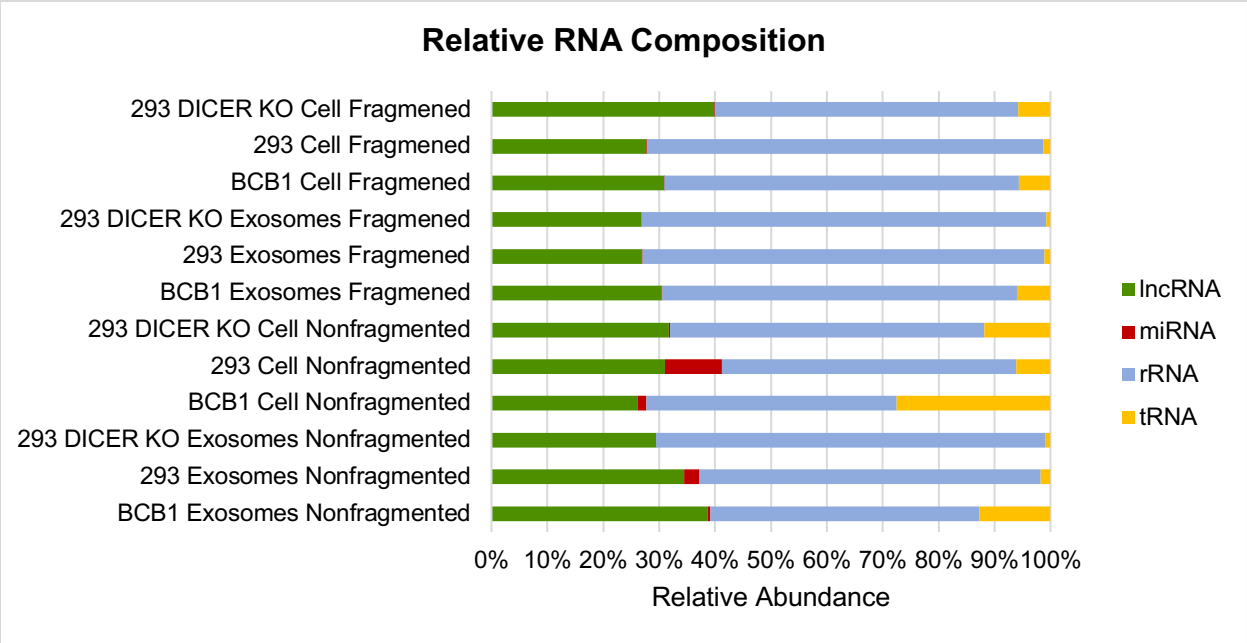


Fig. 7: Average relative RNA composition of exosomal and cellular samples. Clean reads were aligned to known ESTs, lncRNAs, miRNAs, mRNAs, rRNAs, snoRNAs, and tRNAs from the human genome. ESTs are not shown as they accounted for ≥ 90% of the RNA composition in each sample and overwhelmed the graph. mRNAs and snoRNAs are not shown because no positive hits were confirmed in any samples. The average of three biological replicates is shown here.

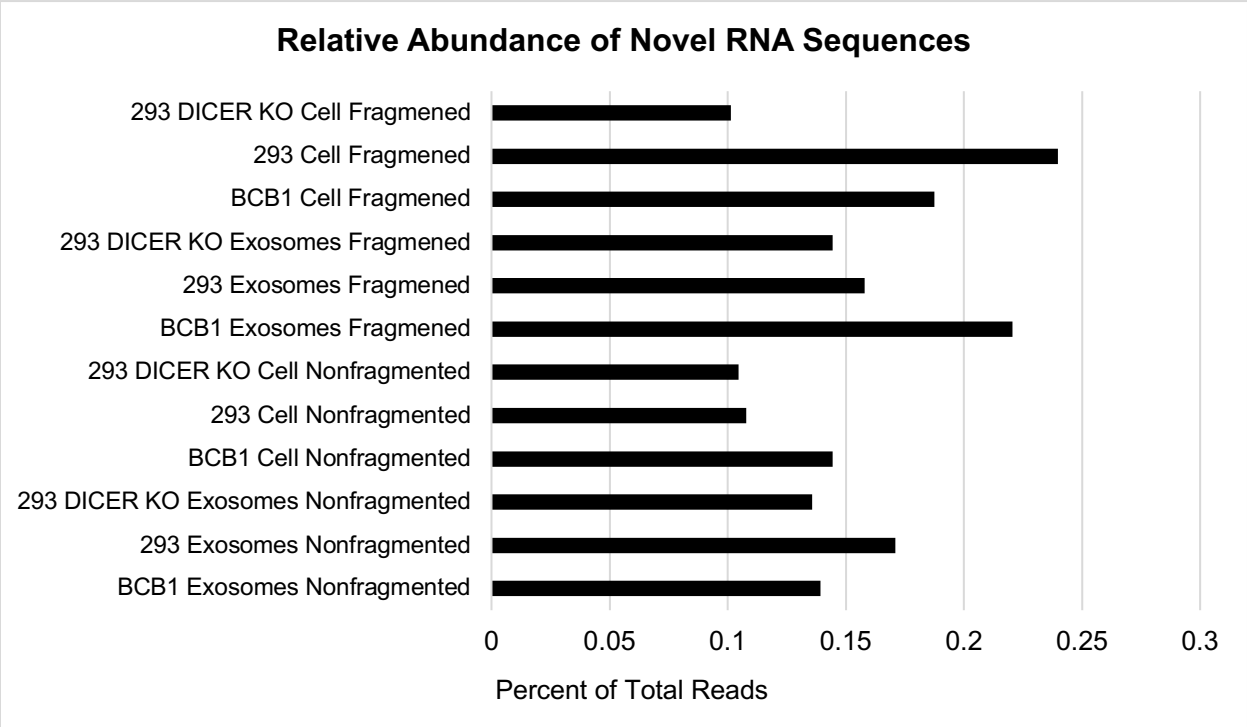


Fig. 8: Average relative abundance of previously unannotated RNA sequences. Clean reads were aligned to known ESTs. The number of reads in each sample that did not align to ESTs are plotted as a function of the total number of reads per sample. The average of three biological replicates is shown here.

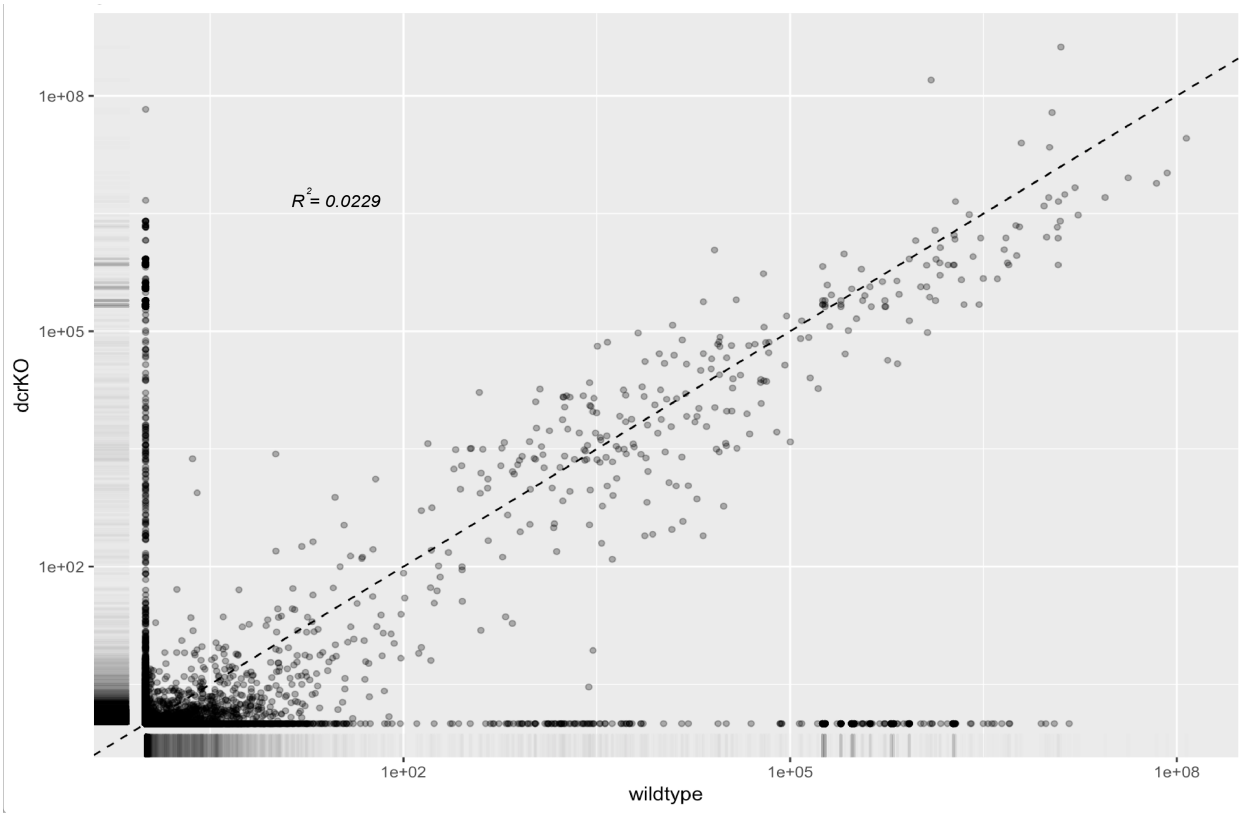


Fig. 9: Scatter plot of RNAs isolated from exosomes from wild-type and Dicer-knockout HEK293 cells reveals differential transcriptome composition. 210 sequences are identified as significantly enriched in wild-type cells. Of these, 81 were mapped to existing RNA annotations in the human genome hg38. The average of three biological sample replicates is shown here. Plot was generated using the R package CummeRbund[7].

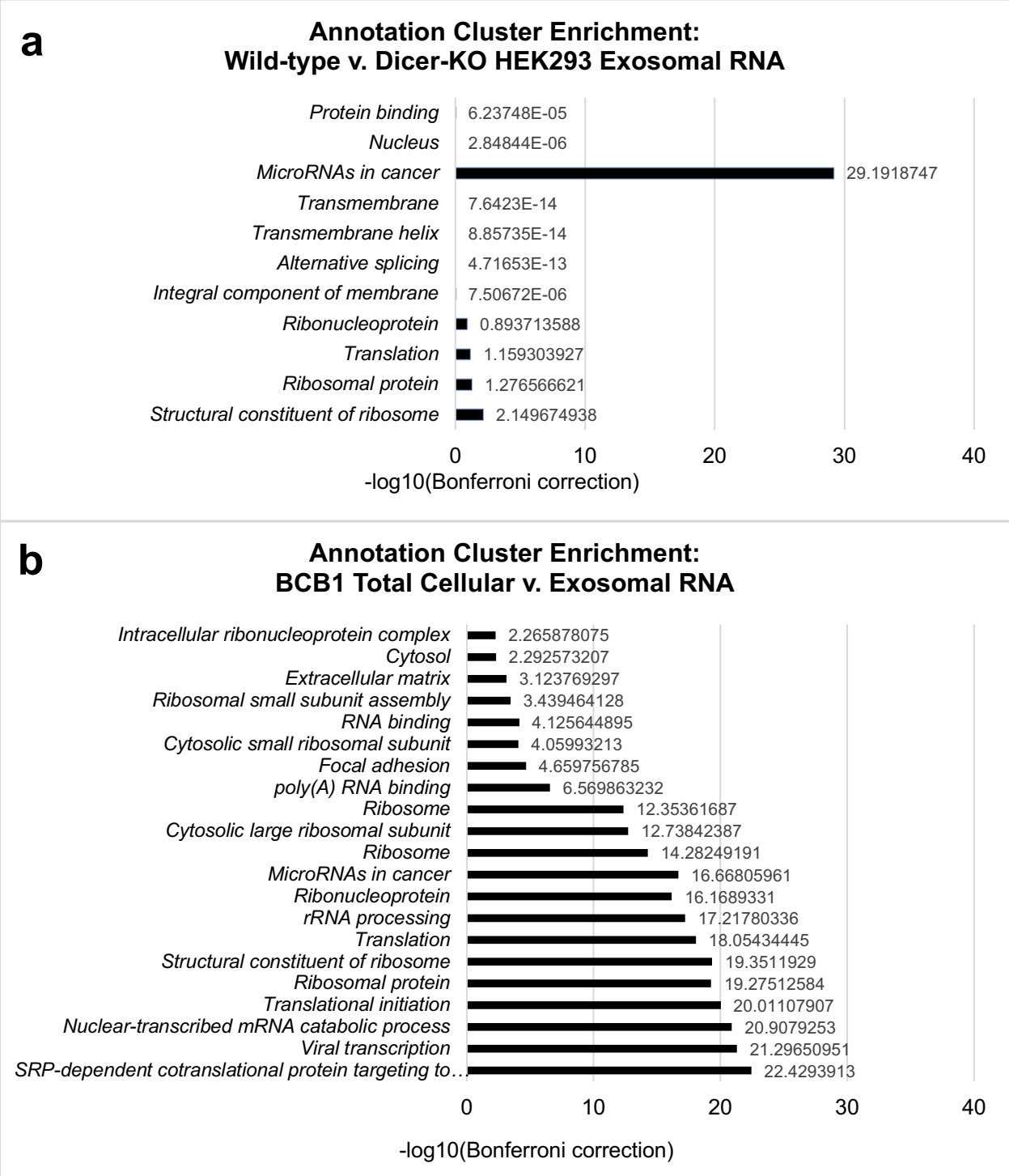
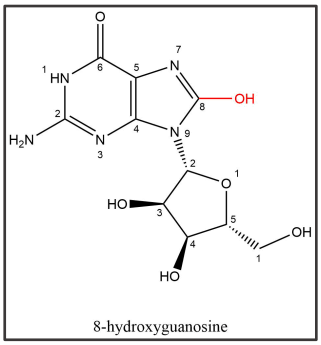


Fig. 10: Functional annotation cluster analysis of sequences that are significantly enriched. Enrichment scores are based on a corrected p-value, the Bonferroni correction. (a) In wild-type exosomes, 11 major functions are revealed. The microRNAs in cancer pathway is most significantly enriched. (b) In BCB1 exosomes, 49 enriched functions were identified. The clusters with a Bonferroni correction ≤ 1.0E-4 are shown. A number of enriched clusters overlap with the wild-type exosomes, including miRNAs in cancer.

Discussion: The results of quantitative LC-MS (Fig. 1) reveal a number of unexpected findings about the abundance of base modifications. Most drastically, an unknown modification with a mass of 299.1 amu is found at high levels in exoRNA, but in negligible levels in total cell RNA. Furthermore, this modification appears in wild-type HEK293 cells at a rate 10 times higher than in Dicer-knockout cells. Previously, this modification was erroneously characterized as ac4Cm (N4-acetyl-2′-O-methylcytidine) based on mass. To validate the authenticity of the modification's identity, a second round of mass spectrometry was performed using a commercial ac4Cm standard (Fig. 2a). Surprisingly, although the precursor ion was the correct mass, the retention time of the experimental peak was much lower than that of the standard peak (22 and 46 minutes, respectively). This demonstrates that the modification found at high levels in wild-type exosomal RNA is not, in fact, N4-acetyl-2′-O-methylcytidine.

To identify the specific RNA species containing this 299.1 amu modification, we size-separated the HEK293 wild-type and Dicer-knockout cellular RNAs and exoRNAs using denaturing polyacrylamide gel electrophoresis (data not shown). Each sample separated into two bands approximately 15-35 nt and 50-75 nt in length; these samples were collected, hydrolyzed, and sent for LC-MS. The results of this analysis indicates that the unknown modification is present on both small RNAs and long RNAs (Fig. 2b). Based on this finding, we propose that the modification is either a mechanism of sorting RNAs into exosomes, or is a result of the RNA in the exosomes being biologically isolated. For example, one possibility based on the mass of the precursor ion is 8-hydroxyguanosine (229.24 amu), a product of oxidative damage



8-hydroxyguanosine

that has been characterized on RNAs, including miRNAs[8]. This damage modification is typically repaired by enzymes such as hOGG1 *in vivo*[9]; however, biological isolation of 8-oxoG modified RNAs via exosomes could serve to size exclude the required repair enzymes and allow for increased stability of the damaged base.

The RNA-seq experiment reads that were aligned to a known human miRNA database have significantly lower mapping rates than expected (Fig. 4). Typical mapping rates for these experiments are somewhere in the range of 40% to 60%. Here, our highest mapping rate was 35.7%. This is the first indication that there may have been an issue with this sequencing run. Additionally, we expect that the Dicer-knockout cells will have a significantly lower mapping rate to the miRNA database than either wild-type cell line; Dicer is responsible for trimming double-stranded RNA to form miRNAs during their processing, and therefore Dicer-KO cells are expected to have decreased levels of miRNAs. However, figures 4 and 5 both demonstrate mapping rates in Dicer-knockout cells on the same order of magnitude as Dicer-wild-type cells. This is the second indication that we experienced an issue during sequencing.

To address the sequencing issues, we examined the read counts of individual miRNA species by hand. Unfortunately, the top 10 most abundant miRNAs in each sample were species which are being studied by other members of the lab. These miRNAs are found at extraordinarily high levels in both exosomes and cells, wild-type and Dicer-knockout species, and across all three replicates of each condition. Therefore, we conclude that the sequencing runs were contaminated with exogeneous RNAs. Figure 6 demonstrates the large discrepancy between the contaminant species and the species of interest, showing how the contaminant species account for > 97% of matched reads. To obtain usable data from these sequencing runs, we capitalized on the presence of a

unique molecular identifier (UMI) in the contaminant species that was not present in the species of interest. Contaminant species were therefore longer than species of interest by a defined amount, and could be removed using a length filter prior to downstream analysis (see Materials and Methods).

For downstream analysis, the clean RNA reads were sent through two different pipelines. Firstly, to characterize the collection of exosomal and cellular RNA in each cell line, we mapped the clean reads from each sequencing run to a database containing a number of RNA subtypes using bedtools (Fig. 7)[10]. The results show that miRNAs are seen at higher relative levels in the non-fragmented samples, which are biased towards small RNAs. rRNA levels are lower than expected across all conditions, largely as a result of rRNA depletion. We also confirm that miRNA levels are significantly reduced in HEK293 Dicer-knockout cells as compared to either wild-type cell line. More specifically, miRNAs are generally enriched in exosomes, consistent with previous findings[11]. Each sample shown in Figure 7 was also aligned to a database of ESTs, a comprehensive list of all previously discovered sequences in the human genome. Less than 1% of reads per sample were not matched to the EST database and are therefore classified as novel reads (Fig. 8).

Secondly, the clean RNA reads were treated using the TopHat and Cufflinks software to align reads to the human genome across splice junctions, assemble the reads, and compare differential expression across conditions. We specifically examined the RNA species isolated from exosomes from wild-type and Dicer-knockout HEK293 cells (Fig. 9). Of the transcriptome species that are found at different levels, 210 are significantly enriched in the wild-type cell line, as determined by the q-value, corrected for the use of three biological replicates. This finding indicates that a portion of these species

are Dicer-associated microRNAs. Additionally, the R^2 value of this plot is 0.0229, indicating a large difference between the wild-type and Dicer-knockout exosomes.

Of the 210 RNA species that are significantly enriched in wild-type exosomes, 81 mapped to previously discovered RNAs in the human genome. We analyzed these 81 transcripts using the DAVID Functional Annotation Cluster Analysis tool (Fig. 10a)[12]. This test indicated 11 major transcript functions that showed significant clustering and enrichment, as determined by the Bonferroni correction value. The minimum number of transcripts required to create a cluster was 3; all clusters contained 3 transcripts, with the exception of alternative splicing (5 transcripts) and microRNAs in cancer (25 transcripts). The microRNAs in cancer pathway is by far the most significantly enriched, with a –log10(Bonferroni correction) of nearly 30.

The same functional clustering analysis was performed using transcripts that were significantly enriched in BCB1 exosomal RNA as compared to BCB1 total cellular RNA (Fig. 10b). 159 significantly enriched RNA species were mapped to the human genome, and 49 significantly enriched clusters were identified from this set of transcripts. Of the 11 enriched wild-type exosomal clusters, 6 are also enriched in BCB1 exosomes, including microRNAs in cancer. In the BCB1 cancerous human cell line, 22 distinct RNA species were grouped into the microRNAs in cancer enriched pathway, the highest of any cluster.

A possible implication of these findings is that small regulatory RNAs, such as miRNAs, are secreted by cells into exosomes *in vivo* and affect gene regulatory activity across distances and in different tissue types. That many miRNAs involved in cancer pathways are significantly enriched in exosomes isolated from wild-type and cancerous human cell lines suggests that exosomes may be involved in cancer communication pathways, specifically.

Conclusions and Summary: The mass spectrometry data reveals interesting insights into the modification landscape of exomiRs. It appears that the unknown modification is found in significantly higher proportions in exosomes than in total cellular RNA. It is also found ubiquitously across exomiRs of many lengths. Though the authenticity of this peak has yet to be confirmed, this data could support the hypothesis that the sorting of small RNAs into exosomes is non-random. Confirming the identity of the modification will also answer outstanding questions about the function of the modification and the mechanism behind its enrichment in exosomal RNA species.

Although the data obtained from RNA-seq experiments was heavily contaminated, removal of the contaminated reads recovered a clean data set that provided many useful characteristics of the exosomal RNA modification landscape. Sequencing data suggests that there is an unequal representation of RNA in exosomal and cellular contents from the same cell line. This supports the hypothesis that RNA sorting into exosomes is nonrandom. It is possible that the unknown modification identified here serves as a biological marker for sorting into exosomes, but future studies are necessary to test this claim.

A popular hypothesis, although still unconfirmed, is that exosomes and their corresponding RNA contents are involved in intercellular communication[11]. Many scientists specifically seek to investigate this claim in the context of the tumor microenvironment. Findings from our RNA-seq differential comparison and functional clustering analysis experiments reveal that RNAs associated with the microRNAs in cancer pathway are (1) significantly enriched in exosomes, as compared to either Dicer-knockout exosomes or total cellular RNA contents, and (2) highly enriched in a transcript clustering analysis, as compared to baseline human genome expression levels. The

microRNA species identified in these clusters have the potential to impact the transcriptome of cells within the tumor microenvironment, possibly leading to altered miRNA levels in receiving cells and, subsequently, altered mRNA target expression levels. Whether this is a ubiquitous mechanism by which cancer cells communicate *in vivo* remains to be confirmed.

Future work seeks to characterize the 299 amu modification presented here, and to identify whether it is part of a sorting mechanism for exoRNAs or whether it is preserved in exoRNAs via biological isolation. Additionally, further experimentation will confirm the effects of exomiRs on recipient cell gene expression levels, specifically those involved in canonical cancer pathways. Finally, the *ex vivo* experiments presented in this paper can be replicated *in vivo* to confirm that trends in exosomal total RNA and miRNA composition remain consistent.

Materials and Methods:

Exosome Isolation: Cells were grown to a density of 800,000 cells/mL in exosome-depleted FBS and centrifuged at 1,000 rpm at 4 °C for 10 min. The supernatant was passed through a 0.22 μm sterile filter, and clarified supernatant and 200 mL sterile PBS were added to a tangential flow chamber for tangential flow filtration under the following conditions: 150 rpm stirring, 20 mL/min flow rate, 1.5 Pr/Pf rate, 50 lmh flux. The flowthrough fraction containing molecules < 250,000 Da was harvested, washed and concentrated. The final volume of clarified supernatant (30 mL) was transferred to a 50 mL conical, and polyethylene glycol was added to a final concentration of 40 mg/mL. This solution was rocked at 4 °C overnight to precipitate exosomes and then centrifuged at

1,200 * g at 4 °C for 1 hr. The pellet was resuspended in 500 μL of PBS and treated with 10 μL of RNAse A (4 mg/mL) to remove exogenous nucleic acids.

RNA Isolation: Exosomal RNAs were isolated using the Total Exosome RNA and Protein Isolation Kit (ThermoFisher) according to the manufacturer's protocols. For total cell RNA isolation, cells grown to a density of 8 x 10^6 cells/mL were washed with PBS and resuspended in 1 mL TRIzol. Chloroform (30% v/v) was added and the solution was centrifuged at 13,300 rpm for 5 min. The aqueous layer was removed and isopropanol (100% v/v) and glycogen (10% v/v) were added prior to a second spin step under identical conditions. The supernatant was removed and RNA pellets were washed with 70% ethanol prior to resuspension in ddH_2O.

RNA-seq Library Preparation: A fraction of total cell RNA from each cell line (500 ng) was rRNA-depleted using the NEBNext rRNA Depletion Kit according to the manufacturer's protocols. Next, a fraction of rRNA depleted total cell RNA and exosomal RNA were fragmented to allow for sequencing of long RNAs during RNA-seq. The total amount of recovered total cell rRNA-depleted RNA and exosomal RNA (500 ng) were each resuspended in 5X RNA fragmentation buffer (1 M Tris, pH 8.0, 2 mM $MgCl_2$) to a final concentration of 40 ng/μL and incubated at 75 °C for 10 min. CIP (20 units, NEB) and NEBuffer 2 (10X, NEB) were added and samples were incubated at 37 °C for 30 min. RNA was isolated using a standard phenol-chloroform extraction prior to 3'-end and 5'-end adapter ligation. Reverse transcription q-PCR was performed with Phusion polymerase (NEB) according to the manufacturer's protocols. Finally, the cDNA PCR products were purified using non-denaturing polyacrylamide gel electrophoresis (10%

APS, 3% TEMED). Bands containing cDNA approximately 120 – 160 nt in length were cut out, extracted in 1 mL ddH_2O, purified using non-binding spin columns, and precipitated using ethanol/sodium acetate prior to RNA-sequencing.

LC-MS Sample Preparation: A fraction of total cell RNA (50 ng) and exosomal RNA (30 ng) from each cell line was hydrolyzed in RNA hydrolysis buffer (50 mM sodium carbonate, pH 9.2, 1 mM EDTA) with PP1 (10 units, NEB), CIP (20 units, NEB), and benzonase (10 units, Sigma-Aldrich). Samples were incubated at 37 °C for 1 hr and then purified on a spin column to remove enzymes. Fresh aliquots of the HEK293 wild-type and Dicer-knockout cellular and exosomal RNAs were then size-separated using denaturing polyacrylamide gel electrophoresis. For each sample, a band of 15-35 nt and 50-75 nt were collected, hydrolyzed, and sent for LC-MS.

Reads annotation: Sequencing reads which resulted from contamination were removed based on a length filter prior to downstream analysis. Clean reads were treated using TopHat and Cufflinks software[7]. Reads were first mapped to the hg38 human reference genome using TopHat splice junction mapper, and then assembled using Cufflinks. Assemblies from different samples were then merged using Cuffmerge, and RNA expression analyses were performed using Cuffdiff. Clean reads produced from this protocol were aligned to the following human sequences using bedtools intersect (parameters: -s -wb)[10]: ESTs, linc RNAs, miRNAs, mRNAs, rRNAs, snoRNAs, and tRNAs (UCSC Genome Browser).

Data analysis: Unless otherwise noted, Excel was used to generate all plots representing sequencing and RNA expression data. Statistical significance for RNA expression is classified as a q value ≤ 0.05. Functional annotation clustering of significant sequences was performed using DAVID 6.8[12]. The minimum number of genes required to create a cluster was set to three.

Acknowledgements: I am grateful to Dr. Dirk Dittmer and Ryan McNamara for developing and executing the protocols for the cell growths, exosome isolation, and RNA isolation. I would also like to thank Leonard Collins, who processed all mass spectroscopy experiments, for his time and flexibility.

References:

1. Wang, W.; Luo, Y.-P. *Journal of Zhejiang University SCIENCE B* 2015, *16*(1), 18–31.

2. Tang, Y. et al.; *Int J Mol Med.* 2017, *40*(3) 834-844.

3. Bhome, R. et al.; *Cancer Letters* 2018, *420*, 228-235.

4. Lin, C.; Li, X.; Zhang, Y.; Guo, Y.; Zhou, J.; Gao, K.; Dai, J.; Hu, G.; Lv, L.; Du, J.; Zhang, Y. *Oncotarget* 2015, *6*(11), 8434–8453.

5. Niu, Y.; Zhao, X.; Wu, Y.-S.; Li, M.-M.; Wang, X.-J.; Yang, Y.-G. *Genomics, Proteomics & Bioinformatics* 2013, *11*(1), 8–17.

6. Sakurai, M.; Ohtsuki, T.; Watanabe, Y.-I.; Watanabe, K. *Nucleic Acids Symposium Series* 2001, *1*(1), 237–238.

7. Trapnell, C.; et al. *Nature Protoc.* 2012, *7*(3), 562–578.

8. Wang, J.X. et. al.; *Mol. Cell.* 2015, *59*(1), 50-61.

9. Bruner, S.D.; Norman, D.P.; Verdine, G.L.; *Nature* 2000, *403*(6772), 859-66.

10. Quinlan, A.R.; Hall, I.M. *Bioinformatics* 2010, *26*(6), 841-42.

11. Wei, Z. et. al.; *Nat. Commun.* 2017, *8*, 1145.

12. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. *Nature Protoc.* 2009, *4*(1), 44-57.